https://doi.org/10.7840/kics.2024.49.5.752

강화학습기반 방향성을 고려한 자율 어뢰 기동 제어

노 지 민◆, 이 현 수*, 박 수 현°, 김 중 헌*, 김 건 형**, 김 승 환**

Directional Autonomous Torpedo Maneuver Control Using Reinforcement Learning

Emily Jimin Roh◆, Hyunsoo Lee*, Soohyun Park°, Joongheon Kim*,
Keonhyung Kim**, Seunghwan Kim**

요 약

본 논문에서는 해양 작전에서의 자율 어뢰의 목표 지점 도달을 위한 어뢰 기동을 최적화하는 방법을 제안한다. 이때 어뢰의 유연한 기동을 위해 다양한 각도로의 방향성을 고려한다. 또한, 실제 해양 환경에서의 지형물로 인한 장애물과 어뢰의 각도 전환 시 발생하는 변침점을 억제하여 어뢰 기동의 효율성을 증대하고자 한다. 본 연구에서는 최대회전각도에 따른 다양한 방향으로 움직이는 에이전트의 행동을 반영한 환경을 정의한다. 이후 Markov Decision Process 기반 강화학습 알고리즘인 Q-Learning을 사용하여 어뢰 기동 전략을 수립한다. 최종적으로 본 알고리즘이 일반적인 Q-Learning 알고리즘에 비교하여 목표 지점까지의 도달 성공률 및 변침점 생성 개수를 통해 해당 알고리즘의 우수성을 입증하고, 실제 해양 환경에서의 적용 가능성을 제시한다.

키워드 : MDP, 강화학습, Q-Learning, 자율 어뢰, 방향성 제어
Key Words : MDP, Reinforcement Learning, Q-Learning, Autonomous Torpedo, Directional Control

ABSTRACT

This paper proposes a method to optimize the autonomous torpedo maneuver path for reaching the target of torpedoes, which are explosive projectile weapons in naval operations. For flexible maneuvering of torpedoes, movement in various directions is considered. Also, the obstacles in the actual marine environment and the minimization of the waypoint that occurs when the angle of the torpedoes is changed considered to increase the efficiency of torpedo maneuvering. Consequently, this study presents the environment that reflects the action of the torpedo in various directions according to the maximum rotation angle. Torpedo maneuver strategy is formulated by applying a Markov Decision Process based reinforcement learning algorithm, Q-Learning. Compared to the general Q-Learning algorithm, the superiority of the proposed algorithm is assessed and its applicability in the actual marine environment, through the success rate of reaching the target point and the number of waypoints.

※ 본 연구는 2022년 정부(방위사업청)의 재원으로 국방기술진흥연구소의 지원을 받아 수행되었습니다. (No. KRIT-CT-22-023, 21-107-D00-012, 잠수함 표적식별 및 교전지원 지능화 기술)
◆ First Author : Korea University Department of Electrical and Computer Engineering, emilyjroh@korea.ac.kr, 학생회원
° Corresponding Author : Sookmyung Women's University Division of Computer Science, soohyun.park@sookmyung.ac.kr, 정회원
* Korea University Department of Electrical and Computer Engineering, hyunsoo@korea.ac.kr, 학생회원; joongheon@korea.ac.kr, 종신회원
** LIGNex1, {gunhyung.kim, seunghwan.kim01}@lignex1.com
논문번호 : KICS2023-10-108-C-RN, Received October 19, 2023; Revised December 28, 2023; Accepted January 21, 2024

Ⅰ. 서 론

해양 작전에서의 어뢰는 함격 공격용 수뢰로 유연한 이동과 파괴력을 갖춘 중요한 무기이다. 특히 자율 어뢰는 기존의 유도 무기와 달리 외부의 인간 조작이나 유도 신호 없이도 목표를 탐지하여 자주적으로 움직이는 무기 시스템이다. 이러한 자율 어뢰는 타격[1] 뿐만 아니라 수중 지형 탐사 및 해역 모니터링[2] 등 여러 임무를 수행할 수 있다. 이렇듯 자율 어뢰는 다양한 상황에서 활용할 수 있는 유망한 시스템이며 인간의 위험을 최소화할 수 있다는 장점을 바탕으로 지속적으로 많이 연구되고 있다[3].

최근 해양 환경에서의 지형 조사를 위한 무인 자율 수중 이동체(Autonomous Underwater Vehicle, AUV)의 기동 전략 수립 연구는 활발히 이루어지고 있다[4]. 이는 사전 데이터 없이도 풍부한 목표 대상물의 특징 밀도에 따라 독립적인 경로 수립을 목표로 한다. 이때 자율 어뢰 기동에 있어 유연한 움직임은 작전의 성공률을 향상시키는 주요 요소이다. 그러나 AUV의 방향성을 고려한 해양 환경에서의 연구는 에이전트의 행동 범위가 넓어짐에 따라 최적의 경로로의 수렴이 어려우므로 지속적인 연구가 필요하다. 따라서 본 논문에서는 어뢰의 최대회전각도에 따른 다양한 방향으로의 이동할 수 있는 에이전트를 고려하여 다양한 방향으로의 유연한 대처가 가능한 자율 어뢰의 기동을 연구한다. 이후 자율 어뢰에 Markov Decision Process (MDP) 기반 강화학습 알고리즘인 Q-Learning[5]을 접목시켜 주어진 해양 환경에서 자율적으로 장애물을 회피하고 목표 지점까지 이동하는 최적의 경로 생성 알고리즘을 제안한다. 이때 자율 어뢰의 효율적인 기동을 위해 각도 전환 시 발생하는 변침점을 최소화하는 방법을 제안한다. 최종적으로 본 알고리즘이 일반적인 Q-Learning 알고리즘에 비교하여 목표 지점까지의 변침점 생성 개수 및 도달 성공률을 통해 해당 알고리즘의 우수성과 실제 해양 환경에서의 적용 가능성을 제시한다.

그림 1은 자율 어뢰 기동을 설명하는 그림으로 해양 환경에서의 장애물을 회피하며 정해진 목표 지점까지 도달하는 시스템 개념도이다.

본 논문은 2장에서 강화학습에 대한 기본 메커니즘을 설명하고 3장에서는 자율 어뢰 기동 전략 최적화를 위한 알고리즘의 설계에 대해 설명한다. 이어 4장에서 본 알고리즘의 성능 평가를 제시하고 5장에서 결론을 맺는다.

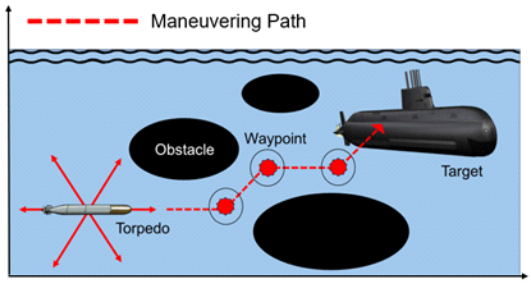


그림 1. 자율 어뢰 기동 개념도
Fig. 1. Overview of Autonomous Torpedo Maneuvering

Ⅱ. 강화학습 개요

2.1 강화학습과 MDP

강화학습(Reinforcement Learning)은 기계학습의 한 종류로 지도학습(Supervised Learning) 및 비지도학습(Unsupervised Learning)과 달리 고정된 데이터셋에 의존적이지 않고 불확실한 환경에서 수집된 경험을 통해 학습된다[6]. 따라서 강화학습은 지도 및 비지도 학습에서는 필수적인 데이터 수집과 레이블 지정 과정을 생략할 수 있는 장점이 있다. 이때 강화학습에서의 경험은 결정을 내리고 행동하는 주체인 에이전트(Agent)가 주어진 특정 환경(Environment)에서의 상호작용을 통한 시행착오 과정을 거쳐 생성된다. 에이전트는 주어진 현재 상태(State)에서 자신의 행동(Action)을 결정하고 환경으로부터 보상(Reward)을 얻어 다음 상태로 넘어가게 된다. 이러한 과정을 반복하여 에이전트는 보상을 통해 스스로 학습하여 최적의 정책(Policy)을 찾는다. 이렇듯 시간의 흐름에 따라 결정을 내리는 문제를 순차적 의사결정(Sequential decision making) 문제[7] 라고 한다. 또한 정책이란 에이전트의 현재 상태를 입력으로 받았을 때 에이전트가 취할 수 있는 행동을 출력하는 함수이다. 에이전트는 자신의 정책에 종속적으로 행동을 취하므로 높은 보상을 받기 위해서는 최적의 정책을 학습하는 것이 중요하다.

강화학습은 순차적 의사결정 문제를 해결하기 위해 활용되며 MDP[8]를 통해 다양한 상태와 행동을 가지는 시스템에서의 문제를 수학적으로 모델링할 수 있다. MDP는 현재 상태가 과거 상태들과 관계없이 오직 직전의 상태에만 의존한다는 마르코프 성질(Markov property)을 기반하여 수식 1과 같이 구성된다.

$$\mathrm{MDP} \equiv (S,\ A,\ P,\ R,\ \gamma) \tag{1}$$

이때 S는 상태, A는 행동, P는 전이 확률 행렬(Transition probability matrix), R은 보상, γ는 감쇠 인자(Discount factor)이다. 이때 상태 집합 (S)은 에이전트인 어뢰가 환경에서 가질 수 있는 모든 상태가 포함된 집합으로 3장에서 자세히 정의한다. 행동 집합 (A)은 에이전트인 어뢰가 해당 환경에서 취할 수 있는 모든 행동을 포함하는 집합이다. 본 논문에서는 에이전트인 어뢰의 행동 집합은 최대회전 각도에 따라 설정될 수 있는 방향 집합에 해당한다. 전이 확률 행렬 (P)는 시간 t에서 특정 상태 S_t에 있을 때, 에이전트가 특정 행동을 취함에 따라 다음의 상태로 전이될 확률을 나타낸다. 보상 (R)은 특정 상태 S_t에서 에이전트가 특정 행동 a_t를 수행하였을 때, 해당하는 보상 값을 출력하는 함수이다. 감쇠 인자 (γ)는 미래의 보상을 현재의 가치로 환산하는 요소로 0에서 1사이의 값으로 설정된다. 감쇠 인자는 에피소드를 진행해가며 계속해서 곱해지는 가중치이므로 가중치의 값이 0에 가까울수록 미래에 대한 보상이 0에 가까워지는 속도가 빨라진다. 즉, γ 값이 0에 가까울수록 현재 상태에 받을 수 있는 보상에 더 집중하게 되며 1에 가까울수록 미래에 얻을 보상에 더 큰 가중치를 두어 행동하는 에이전트가 된다.

강화학습에서의 모든 에이전트는 누적된 보상의 합인 리턴(G_t)을 최대화하는 방향으로 학습이 진행된다. 이때 리턴은 특정 시간 t로부터 미래에 대한 가중치 값이 곱해지며, 감쇠 된 보상의 합으로 수식 2와 같이 계산된다.

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \quad (2)$$

이때 γ는 항상 1보다 작게 설정되게끔 제약을 두어 리턴이 무한대로 발산하는 것을 방지할 수 있다. 만약 γ값이 1보다 크다면 이에 리턴이 무한대로 발산할 수 있으므로 리턴 간 비교가 힘들어지며 학습이 원할 하게 이루어지지 않을 수 있다. 그러므로 γ값을 항상 1보다 작게 설정하여 리턴의 수렴을 보장할 수 있다. 따라서 리턴은 에피소드가 종료될 때 얻는 보상의 합계를 나타내며 에이전트가 특정 시간 동안 얼마나 높은 보상을 얻었는지 측정할 수 있다.

리턴은 정책 함수 (π)에 기반하여 결정된다. 정책 함수란 특정 시간 t에서의 상태 S_t에서 에이전트가 취할 행동 a_t를 결정해주는 함수이며 수식 3과 같이 나타낼 수 있다.

$$\pi(a|s) = P[A_t = a \mid S_t = s] \quad (3)$$

에이전트는 환경 속에서 보상의 값을 통해 학습되어 최종적으로 보상을 최대화할 수 있는 방향으로 학습한다. 이때 정책을 수정해 나가며 최종적으로 리턴의 기댓값을 가장 크게 만드는 정책인 최적 정책 (π^*)을 학습한다. 이때 보상을 가장 높일 수 있는 행동을 선택하기 위해서 상태 별 가치를 계산해야 한다. 가치는 수식 4의 벨만 최적 방정식(Bellman optimality equation)[9]을 통해 계산된다.

$$q_*(s_t|a_t) = R + \gamma \sum_{s' \in S} P_{ss'}^{a} \max\,[\,q_*(s', a')] \quad (4)$$

여기서 $q(s_t|a_t)$는 상태-액션 가치 함수이며 현재 상태 s_t에서 취할 수 있는 a_t의 가치를 평가해주는 함수이다. 일반적으로 Q 가치(Q-Value)로 표현된다. $P_{ss'}^{a}$는 현재 상태 s에서 행동 a를 취하여 다음 상태 s'로 전이할 때 환경으로부터 받는 전이 확률에 해당한다. 이때 벨만 최적 방정식의 max 연산자를 통해 정책을 기반으로 Q 가치를 높이는 가장 높은 행동을 취하게 된다. 따라서 학습이 거듭될 수록 최적의 가치를 가지는 수식 5의 최적 정책을 찾을 수 있다.

$$\pi_* = arg \max_a q_*(s_t|a_t) \quad (5)$$

이러한 메커니즘을 갖는 강화학습은 효과적으로 순차적인 의사결정 문제를 해결할 수 있다. 이때 강화학습의 대표적인 방법으로는 동적 계획법(Dynamic programming)[10]이 있다. 동적 계획법은 벨만 방정식을 반복적으로 사용하여 임의로 초기화되어 있던 가치 값들을 보다 실제적인 값으로 근사하는 방법론이다. 하지만 동적 계획법의 경우 Pseudo-Polynomial[11]의 연산 복잡도를 가지므로 비교적 상태 정보가 단순한 문제만 해결이 가능하며 실제의 환경과 같이 복잡하고 많은 상태 정보를 가지는 문제를 해결할 수 없다. 이에 본 논문에서는 강화학습에서 어뢰의 기동 전략을 효과적으로 학습하기 위한 알고리즘으로 Q-Learning 알고리즘을 적용한다.

2.2 Q-Learning

Q-Learning은 에이전트가 상태-행동 쌍에 대한 가치 함수를 업데이트하는 방식의 학습법으로 환경에서의 탐색 기반으로 학습하여 가장 큰 Q 가치를 가지는 액션을 선택하여 문제를 해결하는 방법이다. Q-Learning은 현재 상태의 가치함수와 다음 상태의 가치함수 사이의 관계를 나타낸 벨만 최적 방정식인 수식 6을 기반으로 학습을 진행한다.

$$Q(s,a) \leftarrow Q(s,a) + \eta\left(R + \gamma max Q(s',a') - Q(s,a)\right) \quad (6)$$

이때 η는 학습률(Learning Rate)로 얼마나 빠른 속도로 학습을 진행할지 결정하는 상수이다. 이를 통해 최종적으로 Q-Learning은 현재 상태 s에서 특정한 행동 a를 취할 때 받을 수 있는 Q 가치를 계산한다. Q 가치란 보상의 기댓값을 출력하는 가치 함수이며 이 값을 기반으로 최적 정책이 결정된다.

Q-Learning은 Off-Policy 학습법을 따른다. Off-Policy 학습법이란 타깃 정책과 행동 정책이 다른 학습법을 뜻한다. 이때 타깃 정책은 학습의 대상이 되는 정책이며 행동 정책은 실제 환경과의 상호작용을 통해 경험을 쌓는 정책이다. Q-Learning에서 행동 정책은 에이전트의 탐험(Exploration)을 포함하기 위한 확률 값 ε이 포함되며 타깃 정책은 단순히 Q 가치가 높은 행동만 선택하는 Greedy 정책을 따른다. 이를 통해 Q-Learning은 불확실한 환경에서도 탐색 기반의 최적 정책으로의 학습이 가능하다.

2.3 자율 수중 이동체에서의 강화학습

최근 해양 환경에서의 군사적 목적 및 지형 탐색을 위한 무인 자율 수중 이동체(Autonomous Underwater Vehicle, AUV)의 연구는 활발히 이루어지고 있다. 특히 AUV의 핵심 기술은 크게 수중 동역학적 모델링이 필수적인 선형/선체기술, 실시간으로 해양 데이터를 수집하는 수중 센서 및 신호처리기술 그리고 자율 어뢰의 전략을 도출하는 항법기술로 이루어진다[12]. 이 중 항법기술은 수중 센서 및 신호처리기술 기반 실시간으로 변화하는 불확실한 환경에서 선형/선체기술의 제약조건을 만족하며 목표 대상물에 따라 독립적인 경로 수립을 목표로 한다.

이러한 항법 기술은 시간에 따른 순차적인 의사결정 문제로 경험 기반의 학습법인 강화학습을 통해 효과적인 전략을 도출할 수 있다. 최근에는 AUV의 궤적 제어[13] 및 도킹 제어[14]를 위한 기동 전략을 심층강화학습(Deep Reinforcement Learning, DRL)을 적용하는 방법이 많이 시도되어지고 있다. 하지만 DRL을 활용한 방법은 실시간성이 중요한 AUV의 추론에 있어 시간이 지연될 수 있으며 불확실한 실제 해양환경과 같이 학습의 정도에 따라 심각한 성능열화의 위험성이 존재한다. 이에 딥러닝 기반 알고리즘 대비 최적해 도출이 가능하며 학습에 따른 성능열화 및 시간을 최소화할 수 있는 MDP 기반의 Search 알고리즘인 Q-Learning을 적용하여 제어 전략을 도출하는 방법도 시도되고 있다[15]. 하지만 대부분의 연구에서 Q-Learning을 적용하는 환경에 있어 방향성이 고려되지 않는 경우가 많으며 이로 인해 실제 환경에서 알고리즘을 적용하여 실험했을 때 시뮬레이션 상에서의 주행과 큰 오차를 초래한다. 이렇듯 실제 수중 이동체의 움직임에는 일정 시간에 기동 가능한 방향의 범위가 존재하며 주어진 일정 거리만 주행 가능하므로 이에 방향성이 고려된 적합한 에이전트와 학습환경에서의 연구가 필수적이다.

따라서 본 논문에서는 방향성이 고려된 에이전트와 환경에서의 자율 어뢰 기동 전략을 시간 지연 및 학습에 따른 성능열화를 최소화할 수 있는 Q-Learning을 적용하여 도출한다. 이에 불확실한 해양 환경에서도 탐험을 통한 최적 정책을 학습한다. 따라서 자율 어뢰가 탐험하면서도 학습된 지식을 토대로 안정적으로 주행하도록 하여 최종적으로 다양한 상황에서도 유연하게 대응하는 최적의 기동 전략을 도출하고자 한다.

Ⅲ. 강화학습 기반 방향성이 고려된 자율 어뢰 기동 제어 전략

3.1 실험 환경

본 논문은 자율 어뢰의 최대회전각도 기반한 방향성을 고려하여 그림 2의 시스템 모델에서의 서로 다른 최대회전각도를 갖는 두 환경에서 실험을 진행하였다. 그림 2의 a의 경우, 최대회전각도가 90도일 때의 환경이며 10×10 정사각형 그리드[16]에서 실험을 진행했다. 그림 2의 b에서의 경우, 최대회전각도가 60도일 때의 방향에 따른 이동거리가 모두 같도록 10×10 정육각형 그리드의 환경에서 실험을 진행했다. 따라서 일반적인 Q-Learning의 환경과 달리 최대회전각도에 따라 일반적인 그리드 월드가 아닌 기동 방향에 따른 기하학적 법칙이 적용된 환경에서 에이전트는 학습하게 된다. 이때 에이전트는 정해진 최대회전각도 내의 이동이 가능하며 따라서 주황색과 녹색에 해당하는 공간으로의 이동만이 가능하다. 또한 실제의 해양 지형을 고려하여 회피해야 하는 장애물을 해양 지역을 센싱한 정보를 기준으로 생성하여 적용하게 된다. 다음은 이와 같은 환경에서의 상태, 행동, 보상함수, 기타 변수 등 강화학습 설계에 관해 설명한다.

3.2 상태

강화학습의 에이전트인 어뢰의 상태 정보는 시간 t에 따라 수식 7과 같이 정의했다.

$$S_t = \{s_t,\ a_t,\ n_t\} \quad (7)$$

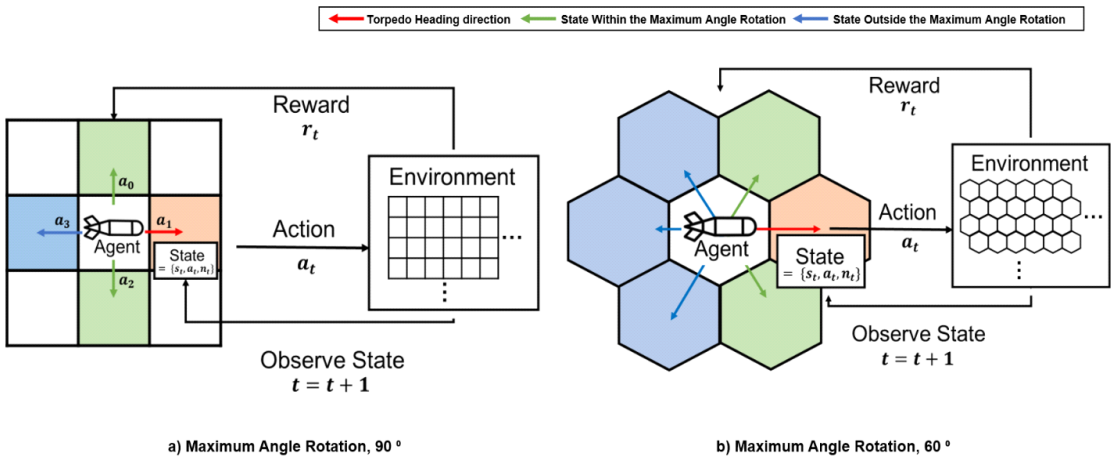


그림 2. 방향성에 따른 시스템 모델
Fig. 2. System Model according to directionality

이때 s_t는 현재 위치 정보, a_t는 어뢰의 행동정보, n_t는 누적 생성된 변침점 개수를 나타낸다. 변침점은 어뢰 기동의 방향이 달라질 때마다 업데이트 되게끔 설정했다. 따라서 n_t는 이 전 상태의 행동인 a_{t-1}과 현재 상태에서의 행동 a_t가 같을 시 그대로 유지되며 같지 않다면 현재의 n_t에 1 증가된다. 이렇게 구성된 상태 정보는 현재의 에이전트가 어떤 상황에 있는지 나타내며 이를 기반으로 에이전트는 행동을 선택하게 된다.

3.3 행동

어뢰는 정해진 최대회전각도에 따라 다른 방향성을 갖고 행동한다. 최대회전각도란 어뢰가 다음 상태로의 기동에서 최대로 회전할 수 있는 각도를 의미한다. 이에 따라서 행동 집합은 그림 2에서의 화살표가 나타내는 방향과 같이 수식 8과 수식 9로 정리할 수 있다.

$$A_{90} = [Up, Right, Down, Left] \tag{8}$$

$$A_{60} = \begin{bmatrix} UpRight, Right, DownRight, \\ DownLeft, Left, UpLeft \end{bmatrix} \tag{9}$$

이와 같이 자율 어뢰는 최대회전각도에 따라 정사각형 및 정육각형으로 이루어져 있는 그리드 환경에서 각 방향으로의 행동이 가능하다. 또한 최대회전각도가 작을수록 기동할 수 있는 방향이 증가한다. 이러한 행동 집합을 가지고 자율 어뢰는 최종적으로 기동 가능한 행동 집합 중에서 가장 효율적인 기동 경로로의 최적의 행동을 선택하게 된다.

3.4 보상함수

어뢰 기동 경로 생성에 있어 효율적인 경로를 생성하기 위해 수식 10과 같이 보상함수를 설계하였다. 이때 변침점의 개수를 고려하여, 어뢰의 방향 변동을 최소화할 수 있도록 하였다. 따라서 보상함수는 (0, α)와 (K, 0)을 지나는 타원의 방정식으로, 변침점의 누적 개수 n_t가 증가할수록 보상은 작아진다.

$$R^{(\alpha, n_t, \theta)} = \sqrt{\alpha^2\left(1 - \frac{{n_t}^2}{K^2}\right)} + \begin{cases} -0.5 & (\theta \le Maximum\ \ Rotation\ \ Angle) \\ -\ 0.8 & (\theta > Maximum\ \ Rotation\ \ Angle) \\ 1 & (Arrive\ \ at\ \ the\ \ Destination\ \ Point) \end{cases} \tag{10}$$

이때 α는 변침점의 생성에 따른 예민도에 해당하는 값으로 0에서 1까지의 값을 가질 수 있다. 예민도는 0에 가까울수록 변침점 생성을 억제한다. K는 어뢰 시스템에서 지정한 고정된 값으로 목적 지점까지의 경로 생성에 있어 생성할 수 있는 최대 변침점의 개수이다. 또한 θ는 에이전트의 회전 각도를 의미한다. 이를 통해 최대회전각도 내에서의 상태에서 보상이 계산된다. 이때 추가적으로 목표지점에 도달 시 + 1 보상, 최대회전각도 밖의 상태 도달 시 - 0.8 보상, 최대회전각도 내의 상태로의 도달 시 - 0.5의 보상을 더했다. 또한 자율 어뢰가 장애물에 도달하게 되면 수식 10에 따라 음수 보상을 받게 된다. 이때 보상함수의 영향을 최대화하기 위해 보상 값은 항상 절댓값이 1을 넘지 않게 설계하였다. 이로써 어뢰는 보상함수를 통해 장애물을 회피하며 변

표 1. 초기 실험 환경 설정
Table 1. Experimental Environment Setup

Notation	Value
환경 사이즈	10 X 10
장애물 수, o	10
최대 변침점 생성 개수, K	15
변침점 생성 예민도, α	0.3
총 에피소드 수, E	2000
감쇠 인자, γ	0.96
학습률, η	0.7
초기 입실론 값, ε_{max}	0.9
에피소드에 따른 입실론 감쇠, ε_{decay}	0.0005

침점 생성을 최소화하고 최대회전각도 밖의 상태로의 이동을 방지하는 최단 경로의 기동 전략을 학습할 수 있다.

3.5 기타 실험 변수 설정

본 어뢰 기동 제어의 초기 실험 환경 설정은 표 1에서와 같다. 학습 초기에는 예측한 가치 값이 불확실하므로 충분히 학습할 수 있도록 ε − *Greedy* 방법[17]을 적용했다. ε − *Greedy* 방법이란 최적의 행동을 선택함과 동시에 일정한 확률로의 무작위 행동을 취하여 새로운 정보를 탐험을 보장하는 방법이다. 이 방법을 통해 강화학습에서의 탐험과 학습한 정보 활용 사이의 균형을 맞출 수 있다. 해당 실험에서는 초기 입실론 값 ε_{max} 를 0.9로 설정하여 첫 에피소드에서 90 %의 확률로 탐색 과정인 무작위 행동을 선택을 하도록 설정했다. 이에 10 %의 확률로는 가장 높은 가치를 갖는 행동을 선택하도록 했다. 이때 에피소드에 따른 입실론 감쇠 값인 ε_{decay}를 0.0005로 설정하여 에피소드를 거듭할수록 경험을 통해 기존에 학습한 정보를 활용할 수 있도록 하였다. 또한 학습률은 0.7로 설정했으며, 감쇠 인자는 0.96으로 설정했다. 이로써 최종적으로 Q-Learning을 통해 어뢰가 장애물을 회피하고, 변침점을 최소화하는 목표 지점까지의 최적해를 찾는다.

Ⅳ. 성능 평가

4.1 최종 보상

강화학습에서 에이전트는 누적된 보상의 합인 리턴을 최대화하는 방향으로 학습이 이루어진다. 그러므로 에피소드에 따라 누적 보상의 합의 변화하는 추이를 통해 해당 에이전트의 성능을 확인할 수 있다. 그림 3은 에피소드가 진행됨에 따라 강화학습 기반 제안 알고리

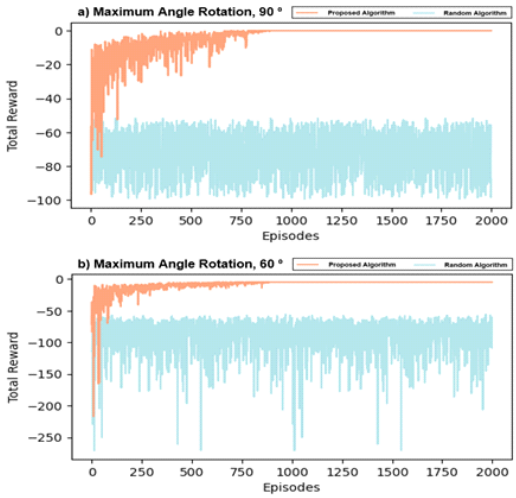


그림 3. 에피소드 진행에 따른 누적 보상 값
Fig. 3. Total Reward Value according to Episodes

즘 및 무작위 알고리즘에서의 누적 보상 추이를 나타낸 그래프이다. 이때 a는 최대회전각도가 90도 일 때, b는 60도 일 때의 결과이다. 최대회전각도에 따른 두 제안 알고리즘 모두 해당 문제에서 무작위 하게 행동을 선택하는 알고리즘에서 보다 높은 보상 값을 가지며, 에피소드가 거듭됨에 따라 누적 보상 값이 증가하여 최종적으로 수렴됨을 확인할 수 있다. 또한 보상 값의 변동 양상 또한 무작위 알고리즘에 비해 적은 것을 통해 제안한 알고리즘에서의 성능이 안정적임을 확인할 수 있다.

4.2 변침점 생성

그림 4는 일반적인 Q-Learning과 변침점 최소화를 위해 본 논문에서 설계한 보상함수를 적용한 제안 알고리즘에서의 목표 지점 도달 까지의 변침점 생성 개수를 비교한 그래프이다. 이때 환경에서의 장애물 개수를 환경 크기로 나눈 값인 환경 복잡도를 증가시키며 관찰했다.

최종 변침점 수는 각 알고리즘을 10회 반복했을 때 평균적으로 생성되는 변침점 개수로 평가했다. 이를 통해 어뢰 기동 전략의 효율성을 평가할 수 있다. 그림 4의 a는 최대회전각도가 90도, b는 최대회전각도가 60도일 때의 변침점 생성 결과이다. 두 방향성에 따른 결과에서 일반적인 Q-Learning 알고리즘에서 보다 제안 알고리즘에서 변침점 생성을 더욱 억제할 수 있는 것을 확인할 수 있다. 억제의 정도는 환경 복잡도가 증가할수록 차이가 더욱 뚜렷해지는 것을 볼 수 있다. 이때 일반적인 Q-Learning 알고리즘에서 보다 a 에서는 약 61.4%가량, b에서는 약 64.7%가량 변침점 생성을 절약

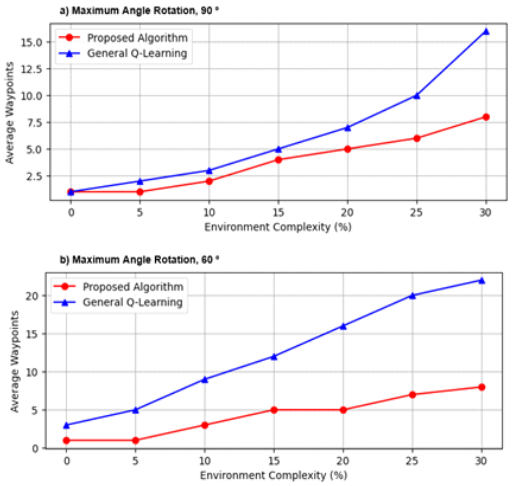


그림 4. 환경 복잡도에 따른 변침점 생성 결과
Fig. 4. Number of Generated Waypoint according to Environment Complexity

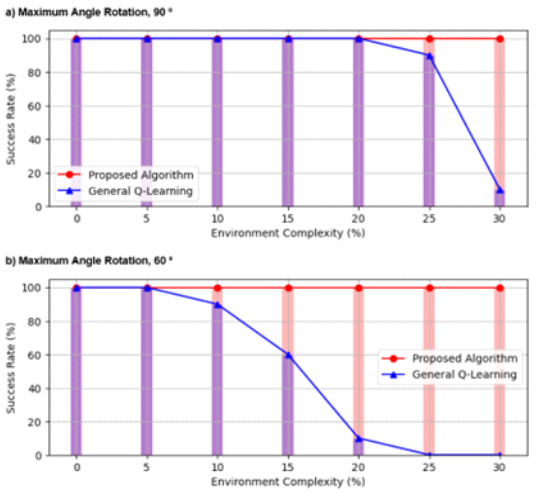


그림 5. 환경 복잡도에 따른 어뢰 기동 성공률
Fig. 5. Torpedo Maneuver Success Rate according to Environment Complexity

할 수 있다. 따라서 제안 알고리즘이 일반적인 Q-Learning 알고리즘에서 보다 목표 지점 도달까지의 방향 회전에 따른 시간적, 비용적 측면에서 매우 효율적인 기동 전략을 도출함을 알 수 있다.

4.3 목표 지점 도달 성공률

그림 5는 제안 알고리즘과 일반적인 Q-Learning 알고리즘의 목표 지점 도달 성공률을 나타낸 그래프이다. 이때 성공 기준은 어뢰 기동 시 제약 조건을 충족하면서 목표 지점까지의 도달이다. 어뢰의 제약 조건은 최대회전각도 밖의 상태로의 이동 금지와 어뢰 시스템에서 생성 가능한 최대 변침점 개수 K 이하의 누적 변침점 개수 유지이다. 그림 5의 a는 최대회전각도가 90도, b는 60도일 때의 실험 결과이며 두 방향성의 제안 알고리즘 모두 환경 복잡도의 증가에도 항상 100 %의 성공률을 유지하는 것을 확인할 수 있다. 반면 일반적인 Q-Learning의 경우 환경 복잡도의 증가에 따라 성공률 또한 현저히 낮아지는 것을 확인할 수 있다. 이는 최대회전각도 밖의 상태로의 이동과 누적 생성된 변침점의 개수 초과로 인한 결과로, 최대회전각도가 60도 일때의 결과가 90도 일 때보다 성공률이 더욱 떨어지는 것을 볼 수 있다. 이는 어뢰의 방향성 증가에 제안 알고리즘의 성능이 더욱 우위에 있음을 알 수 있다. 이를 통해 제안 알고리즘은 환경 복잡도의 변화에도 항상 높은 성공률로 최적의 어뢰 기동 전략을 도출함을 확인할 수 있다.

4.4 각 알고리즘에 따른 어뢰 기동 제어

그림 6은 방향성에 따른 제안 알고리즘과 일반적인 Q-Learning 알고리즘에서의 실험을 시각화한 결과이다. 실험은 환경 복잡도가 10일 때의 환경에서 생성된 어뢰 기동으로 각 알고리즘에 따라 도출된 결과를 통해 보다 객관적으로 본 알고리즘의 성능을 평가할 수 있다. 최대회전각도에 따라 좌측은 90도, 우측은 60도의 두 가지 환경에서 도출된 결과이며, 그림 6의 상단 그래프는 제안 알고리즘, 하단 그래프는 일반적인 Q-Learning 알고리즘에서의 어뢰 기동 결과이다. 그림 6의 a의 경우 제안 알고리즘에서는 최종적인 어뢰 기동에서 총 2개의 변침점이 생성되었고 일반적인 Q-Learning 알고리즘에서는 총 3개의 변침점이 생성되었다. 또한 그림 6의 b의 경우 제안 알고리즘에서 총 2개의 변침점이, 일반적인 Q-Learning 알고리즘에서 총 7개의 변침점이 생성되었다. 이는 제안 알고리즘에서 가장 가까운 거리와 동시에 변침점을 최소화하는 어뢰 기동 전략을 도출함을 알 수 있다. 이를 통해 어뢰 기동 소요 시간을 단축함과 동시에 자원 사용을 줄일 수 있음을 보다 객관적으로 확인할 수 있다.

V. 결 론

본 논문은 불확실한 해양 환경에서의 방향성을 고려한 자율 어뢰의 목표 지점 도달을 위한 강화학습 기반 어뢰 기동 최적화 방법을 제안한다. 이때 실제 해양 환경에서의 장애물과, 어뢰의 효율적인 기동을 위한 변침

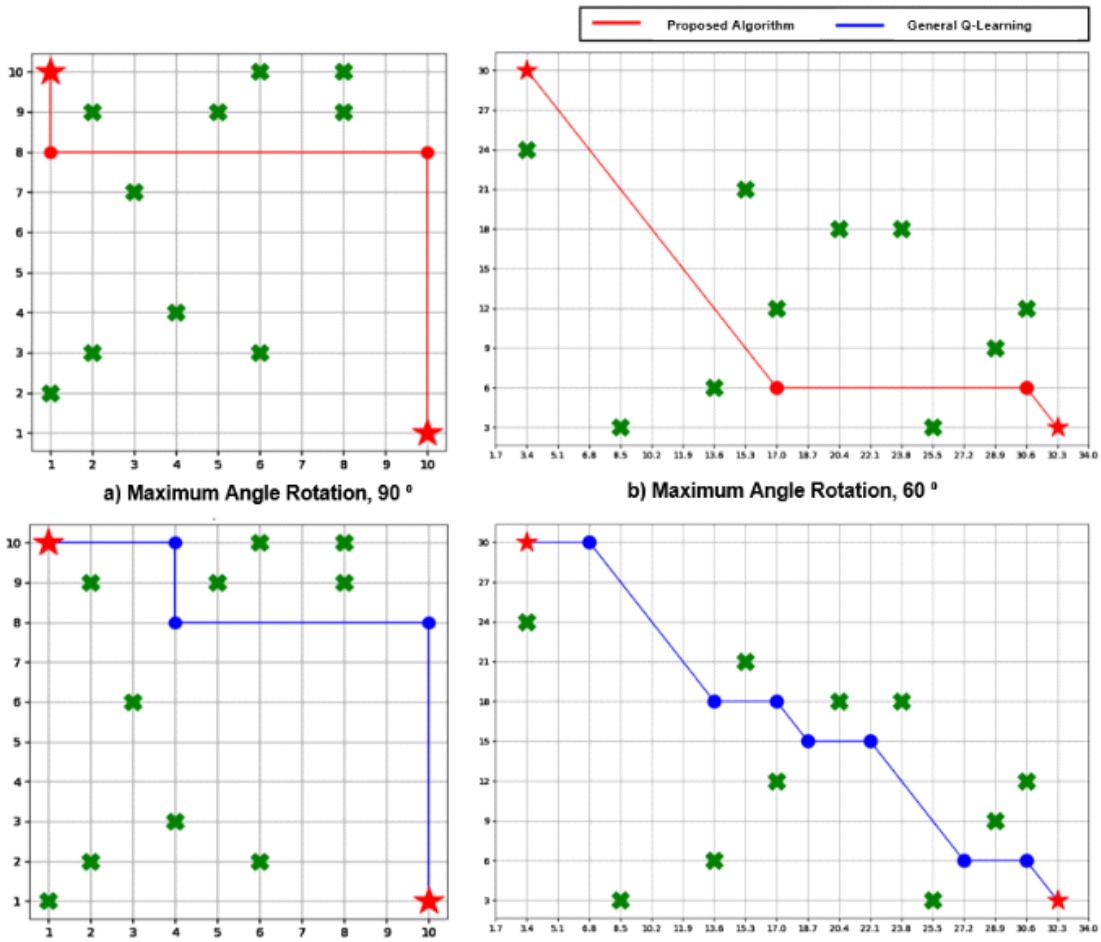


그림 6. 각 알고리즘에 따른 어뢰 기동 결과
Fig. 6. Torpedo Maneuver Result (Proposed Algorithm, General Q-Learning Algorithm)

점 생성 최소화를 함께 고려하였다. 성능 평가를 통해 여러 방향성을 가지는 문제에서 제안한 강화학습 기반 어뢰 기동 알고리즘이 일반적인 Q-Learning 알고리즘과 비교하여 우수한 성능과 안정적인 성공률을 보이며, 변침점 최소화에도 효과적임을 증명하였다. 강화학습은 어뢰가 기동할 수 있는 많은 경우의 수를 따지는 경험 기반 최적의 기동 전략을 도출하는 방법으로 실제 해양에서는 일정 주기마다 업데이트 되는 해양 환경에 대하여 본 알고리즘에 따라 지속적으로 학습하는 과정을 반복하여 더욱 고도화된 기동 전략을 도출할 수 있다. 따라서 본 알고리즘을 통해 실제 해양에서도 어뢰의 자율성 및 효율성을 증대하는 전략을 도출하며 이를 토대로 적용할 수 있을 것으로 기대된다.

References

[1] J. M. Pak, et al., “Target search method for a torpedo to the evading ship using fuzzy inference,” in *Proc. 2009 Int. Joint Conf. ICCAS-SICE,* pp. 5279-5284, Fukuoka, Japan, Aug. 2009. (https://ieeexplore.ieee.org/document/5333400)

[2] L. C. Ignacio, et al., “Optimized design of an autonomous underwater vehicle, for exploration in the Caribbean Sea,” *Ocean Eng.*, vol. 187, no. 17, p. 106184, Sep. 2019. (https://doi.org/10.1016/j.oceaneng.2019.106184)

[3] D. Wang, et al., “Hyperparameter optimization

for the LSTM method of AUV model identification based on Q-Learning," *J. Marine Sci. and Eng.*, vol. 10, no. 8, p. 1002, Jul. 2022. (https://doi.org/10.3390/jmse10081002)

[4] Y. Zhang, et al., "An online path planning algorithm for autonomous marine geomorphological surveys based on AUV," *Eng. Appl. Artificial Intell.*, vol. 118, p. 105548, Feb. 2023. (https://doi.org/10.1016/j.engappai.2022.105548)

[5] C. J. Watkins and P. Dayan, "Q-Learning," *Mach. Learn.*, vol. 8, pp. 279-292, May 1992. (https://doi.org/10.1007/BF00992698)

[6] R. S. Sutton, et al., "Reinforcement learning: An introduction," *Robotica*, vol. 17, no. 2, pp. 229-235, Mar. 1999. (https://doi.org/10.1017/S0263574799271172)

[7] S. Liu, et al., "Understanding sequential decisions via inverse reinforcement learning," in *Proc. IEEE 14th Int. Conf. Mobile Data Manag.*, vol. 1, pp. 177-186, Milan, Italy, Jul. 2013. (https://doi.org/10.1109/MDM.2013.28)

[8] M. Van Otterlo and M. Wiering, "Reinforcement learning and Markov decision processes," *Reinforcement learning: State-of-the-art,* vol. 12, no. 1, pp. 3-42, Mar. 2012. (https://doi.org/10.1007/978-3-642-27645-3_1)

[9] R. Bellman, "Dynamic programming," *SCIENCE*, vol. 153, no. 3731, pp. 34-37, Jul. 1966. (https://doi.org/10.1126/science.153.3731.34)

[10] D. Geiger, et al., "Dynamic programming for detecting, tracking, and matching deformable contours," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 17, no. 3, pp. 294-302, Mar. 1995. (https://doi.org/10.1109/34.368194)

[11] S. A. Cook, "The complexity of theorem-proving procedures," in *Logic, Automata, and Computational Complexity: The Works of Stephen A. Cook*, vol. 1, no. 1, pp. 143-152, May 2023. (https://dl.acm.org/doi/abs/10.1145/3588287.3588297)

[12] D. R. Blidberg, "The development of autonomous underwater vehicles (AUV); A brief summary," *IEEE ICRA*, vol. 4, no. 1, pp. 122-129, Seoul, Korea, May 2001.

[13] Y. Fang, et al., "AUV position tracking and trajectory control based on fast-deployed deep reinforcement learning method," *Ocean Eng.,* vol. 245, no. 1, p. 110452, Feb. 2022. (https://doi.org/10.1016/j.oceaneng.2021.110452)

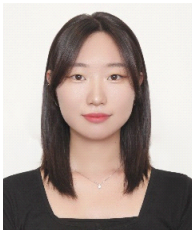
[14] E. Anderlini, G. G. Parker, et al., "Docking control of an autonomous underwater vehicle using reinforcement learning," *Applied Sci.*, vol. 9, no. 17, p. 3456, Aug. 2019. (https://doi.org/10.3390/app9173456)

[15] G. S. Kim, et al., "Utilizing reinforcement learning for enhanced efficiency in fighter airport scheduling: A case study of Osan Air Base," in *Proc. KICS Summer Conference 2023,* vol. 157, pp. 867-868, Jeju, Korea, Jun. 2023.

[16] M. A. Mueller, "Reinforcement learning: MDP applied to autonomous navigation," *Machine Learning and Applications: An Int. J.*, vol. 4, no. 4, pp. 1-10, Dec. 2017. (https://doi.org/10.5121/MLAIJ.2017.4401)

[17] M. Kaloev and G. Krastev, "Experiments focused on exploration in deep reinforcement learning," in *Proc. IEEE ISMSIT*, pp. 351-355, Ankara, Turkey, Nov. 2021. (https://doi.org/10.1109/ismsit52890.2021.9604690)

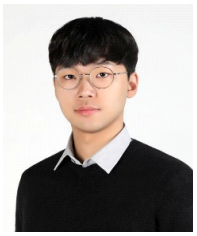
노 지 민 (Emily Jimin Roh)



2024년 2월 : 세종대학교 지능기전공학부 무인이동체공학전공 졸업 (공학사)
2024년 3월~현재 : 고려대학교 전기전자공학과 석박사통합과정

<관심분야> Reinforcement Learning, Autonomous Mobility, Quantum Machine Learning
[ORCID:0009-0008-0013-6342]

이 현 수 (Hyunsoo Lee)



2021년 2월 : 숭실대학교 전자정보공학부 졸업 (공학사)
2021년 3월~현재 : 고려대학교 전기전자공학과 석박사통합과정
<관심분야> Reinforcement Learning, Electronic Engineering, Communication Engineering
[ORCID:0000-0003-1113-9019]

박 수 현 (Soohyun Park)

2019년 2월 : 중앙대학교 컴퓨터공학과 졸업 (공학사)
2023년 8월 : 고려대학교 전기전자공학과 졸업 (공학박사)
2023년 9월~2024년 2월 : 고려대학교 정보통신기술연구소 박사후연구원
2024년 3월~현재 : 숙명여자대학교 소프트웨어학과 조교수
<관심분야> Deep Learning Theory, Network/Mobility Applications, Quantum Machine Learning, AI-based Autonomous Control
[ORCID:0000-0002-6556-9746]

김 중 헌 (Joongheon Kim)

2004년 2월 : 고려대학교 컴퓨터학과 졸업 (이학사)
2006년 2월 : 고려대학교 컴퓨터학과 석사
2014년 8월 : University of Southern California Computer Science 박사
2016년 3월~2019년 8월 : 중앙대학교 소프트웨어대학 조교수
2019년 9월~현재 : 고려대학교 전기전자공학부 부교수
<관심분야> Stochastic Optimization, Mobility, Reinforcement Learning, Quantum
[ORCID:0000-0003-2126-768X]

김 건 형 (Keonhyung Kim)

2009년 2월 : 한양대학교 전자컴퓨터공학 졸업 (공학사)
2007년 12월~2010년 3월 : 삼성탈레스
2010년 4월~2012년 2월 : LG이노텍
2012년 2월~현재 : LIG넥스원
<관심분야> 수중체계, 어뢰기동, 해양
[ORCID:0009-0005-6480-0157]

김 승 환 (Seunghwan Kim)

2009년 2월 : 숭실대학교 정보통신전자공학 졸업 (공학사)
2011년 2월 : 숭실대학원 정보통신공학 석사졸업
2011년 1월~2014년 6월 : 산엔지니어링
2015년 1월~현재 : LIG넥스원
<관심분야> 수중체계, 어뢰기동, 해양
[ORCID:0000-0002-5841-1879]